ANALYSIS METHODS FOR THE QUALITY CONTROL OF OBSERVATIONS

A C LORENC

Meteorological Office UK

1. INTRODUCTION

A major problem for operational meteorological data analysis schemes is the quality control of the observational data. In an intercomparison of three analysis systems, using identical data, Hollingsworth et al (1985) found that in most of the regions of large analysis difference chosen for case studies there were differences in the quality control. In that study we did not demonstrate that differences in quality control decisions caused the analysis differences; in some cases differences in the forecast background field may have caused both. However it was clear that there was scope for improvement in the quality control methods, and that this would have reduced differences between the analyses.

Many analysis systems use the optimum interpolation (OI) analysis method; this can be readily adapted to quality control observations. Methods of doing this are discussed in section 2, with examples taken from the ECMWF analysis scheme (Lorenc 1981). However it is instructive to reconsider the theoretical basis for this, so in section 3 I present some simple studies using a Bayesian approach. These suggest an extension of the methods generally in use for specifying reliability.

If statistically based methods are to be better than older empirically tuned methods, it is vital that the statistics used are appropriate. Methods of collecting these are discussed in section 4.

In section 5 some alternative approaches are discussed; an adaptive Baysian approach (Purser 1984), generalized cross validation (Wahba and Wendelberger 1980), the treatment of mean errors, and interactive human intervention. These methods are appropriate for some of the modern indirect observing systems which can have systematically varying and correlated errors.

In section 6 I summarise, and speculate about longer term developments. More theoretical aspects of the equivalence of OI, the Bayesian approach and variational methods, and of the effect of finite resolution and spectral truncation, are dealt with in appendices.

2. QUALITY CONTROL USING OPTIMUM INTERPOLATION

2.1 Definitions and discussion

It is important that I should start by defining and making clear the difference between the terms quality control, data selection, and observational error specification. A major part of the effort of an operational analysis system is expended in choosing which data to leave out. Two types of data need to be identified; those which are grossly incorrect or misleading (quality control), and those which carry little extra information over others which are being used and which can therefore be disregarded to save time (data selection). Because all data have errors of observation or representativeness, and because of intrinsic or explicit assumptions about the smoothness of fields, the analysis has to be a compromise between the various selected observed values and the background field; the compromise is specified in OI by the observational errors, which determine the relative weights.

The criterion for data selection implied by the above definition is appropriate to the final production of analysed fields; it is not appropriate to quality control which uses the redundant information between observations. Thus for the analysed fields it is appropriate to pre-select a reduced network of observations, for instance by using a pre-determined station list, or by combining nearby observation into 'super-obs', or by choosing to use either height or temperature but not both from radiosonde soundings. This pre-selection is however removing much of the redundant information necessary for quality control, and if it must be done for practical reasons it should be combined with a preliminary quality control which uses the information to be left out.

The definition of observational error makes clear its dependence on the smoothness and resolution of the analysis; error is defined as deviation from a 'truth' which only resolves certain scales of atmospheric motion. Thus it is appropriate to assign an error of about 5 m/s to an aircraft wind report when doing a large scale analysis, even though its real error, as measured by the next aircraft along the same flight path, might be only about 2 m/s. This separation of scales should not affect the quality control criterion derived below, since we will be comparing the deviation of observation from analysis with the sum of analysis and observational error variances; our estimate of analysis error uses the same concept of 'truth' and is correspondingly smaller. However in extreme cases our definition of 'truth' can affect quality control. For instance a wind observation from a ship in a fjord would be correct for an analysis which resolves the fjord, otherwise its error distribution would be so far from normal that it should be rejected, indeed it is difficult to define what the smooth analysis should be in this case. (This is discussed further in Appendix 3).

It is important to realise that the observational error applies to good observations which have passed the quality control checks; it is not a measure of reliability. In practice reducing the specified observational error of a particular type of observation causes more observations to be rejected by a quality control algorithm such as that in section 2.2. Ways of specifying reliability are discussed in section 3.

2.2 Derivation of equations

A derivation of the basic OI analysis equations, and their application to quality control, was set out in a paper describing the ECMWF analysis (Lorenc 1981). For ease of reference I reproduce the relevant parts here, retaining the same notation and equation numbers, before going on to further amplify the discussion.

Notation and basic method

The statistical techniques used are independent of the actual variables observed or interpolated, so I use in this section a notation which does not explicitly differentiate between them, allowing subscripts to range as appropriate over all observed or analysed values whatever their position, level, or variable type. Thus B_i is any observed datum selected for the analysis, and A_k any analysed value within the analysis volume.

For all observed or analysed value I assume the existence of predicted (first-guess) values P_i, P_k and 'true' values T_i, T_k, the last being the value we wish to estimate in the analysis. Note that T is not necessarily the actual true value, since we do not wish to analyse atmospheric features below a certain scale. Deviations from this 'true' value are denoted by lower case letters:-

$$a = A - T \tag{1a}$$

$$b = B - T \tag{1b}$$

$$p = P - T \tag{1c}$$

All analysed, observed or predicted values have associated error estimates E defined by

$$E^a = \langle a^2 \rangle^{1/2} \tag{1d}$$

$$E^o = \langle b^2 \rangle^{1/2} \tag{1e}$$

$$E^p = \langle p^2 \rangle^{1/2} \tag{1f}$$

where $\langle\ \rangle$ indicates an average over a large ensemble of similar realisations. It is convenient to derive equations in dimensionless form, and to have symbols for deviations from the prediction, so I define

$$\alpha = a/E^a \tag{1g}$$

$$\beta = b/E^p \tag{1h}$$

$$\pi = p/E^p \tag{1i}$$

$$q = (B - P)/E^p \tag{1j}$$

$$r = (A - P)/E^p \tag{1k}$$

$$\epsilon^o = E^o/E^p \tag{1l}$$

$$\epsilon^a = E^a/E^p \tag{1m}$$

All the above take subscripts i (or j) ranging over all observed values, or k ranging over all analysed values, whatever their position level or variable.

The basis of the statistical interpolation method is that the analysed deviation from the prediction is given by a linear combination of N observed deviations:-

$$r_k = \sum_{i=1}^{N} w_{ik}\, q_i \tag{2}$$

with the weights (w) determined so as to minimize the estimated analysis error E^a_k.

Substituting (1) in (2) gives

$$\alpha_k \epsilon_k^a = \pi_k + \sum_{i=1}^{N} w_{ik} (\beta_i \epsilon_i^o - \pi_i) \qquad (3)$$

Squaring (3) and taking the ensemble average gives

$$(\epsilon_k^a)^2 = 1 + 2 \sum_{i=1}^{N} w_{ik} (\langle \pi_k \beta_i \rangle \epsilon_i^o - \langle \pi_k \pi_i \rangle)$$

$$+ \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ik} (\langle \pi_i \pi_j \rangle + \epsilon_i^o \langle \beta_i \beta_j \rangle \epsilon_j^o$$

$$- \epsilon_i^o \langle \beta_i \pi_j \rangle - \langle \pi_i \beta_j \rangle \epsilon_j^o) w_{jk} \qquad (4)$$

These summations are conveniently manipulated using a vector and matrix notation, so I define

$$\mathbf{W}_k = [w_{ik}] \qquad (5a)$$

$$\mathbf{P}_k = [\langle \pi_k \pi_i \rangle - \langle \pi_k \beta_i \rangle \epsilon_i^o] \qquad (5b)$$

$$\mathbf{q} = [q_i] \qquad (5c)$$

$$\mathbf{M} = [[\langle \pi_i \pi_j \rangle + \epsilon^o \langle \beta_i \beta_j \rangle \epsilon_j^o \qquad (5d)$$

$$- \epsilon_i^o \langle \beta_i \pi_j \rangle - \langle \pi_i \beta_j \rangle \epsilon_j^o]]$$

(2) and (4) then become

$$\mathbf{r}_k = \mathbf{W}_k^T \mathbf{q} \qquad (6)$$

$$(\epsilon_k^a)^2 = 1 - 2 \mathbf{W}^T \mathbf{P}_k + \mathbf{W}_k^T \mathbf{M} \mathbf{W}_k \qquad (7)$$

I can now proceed to the derivation of the equation for the 'optimum' weights, which minimize E^a. Since the ensemble average $\langle\ \rangle$ is assumed to be over a large number of similar realizations with the same estimated errors E, this is equivalent to minimizing the normalized error variance given by (4) or (7). By equating $\partial(\epsilon^a)^2/\partial w_{ik}$ to zero for i = 1, N we get a set of linear equations for the weights which give:

$$\mathbf{W}_k = \mathbf{M}^{-1} \mathbf{P}_k \qquad (8)$$

The analysed value and estimated error corresponding to these weights are:-

$$r_k = \mathbf{P}_k^T \mathbf{M}^{-1} \mathbf{q} \qquad (9)$$

$$(\epsilon_k^a)^2 = 1 - \mathbf{P}_k^T \mathbf{M}^{-1} \mathbf{P}_k \qquad (10)$$

Since M^{-1} and q are independent of the point being analysed it is convenient to evaluate their product once only, to give a vector of analysis coefficients c. Thus for the grid-point analysis the weights w_k are not explicitly calculated, instead (9) becomes

$$\underset{\sim}{c} = \underset{\sim}{M}^{-1} \underset{\sim}{q} \qquad (11)$$

$$r_k = \underset{\sim}{c}^T \underset{\sim}{p}_k \qquad (12)$$

It is usual to call terms such as $\langle \pi_i \pi_j \rangle$ error correlations and terms such as $\langle p_i p_j \rangle$ covariances, although this is only true if biases such as $\langle p_i \rangle$ are zero. This is not strictly necessary for the above derivation, but if biases are non-zero (2) is not the best interpolation equation. I shall assume the biases to be zero. It is also usual to neglect correlations between prediction error and observation error $\langle \pi_i \beta_j \rangle$ (eg Bergman 1979). If an observation type is to be used for which these terms are known to be non-zero then their inclusion is straightforward.

Observation check

The final check on each datum is to compare q_k with an interpolated value r_k using the data selected for the analysis volume. Hence it is appropriate when deriving the equations for this interpolation to minimize the expected variance of the difference between these, rather than the deviation from the true value. Thus instead of (7) we minimize

$$\langle (r_k - q_k)^2 \rangle = (\epsilon_k^o)^2 + 1 - 2\, \underset{\sim}{w}_k^T \underset{\sim}{m}_k + \underset{\sim}{w}_k^T \underset{\sim}{M}\, \underset{\sim}{w}_k \qquad (18)$$

If the datum being checked is also used for the interpolation then $\underset{\sim}{m}_k$ is a column of $\underset{\sim}{M}$ and minimizing (18) leads to the trivial result

$$\underset{\sim}{w}_k = \underset{\sim}{d}_k \qquad (19)$$

where $\underset{\sim}{d}_k$ is defined as a vector whose k'th element is one and other elements are zero. Since we are trying to interpolate a value for a datum including its observational error, the best value is naturally the datum itself.

What we must do is minimize (18) subject to constraints that certain data (datum k and perhaps other data already rejected) are given zero weight).

If we let l_m (m = 1 to n) be a list of these data, the constraints can be written

$$\underset{\sim}{d}_l^T \underset{\sim}{w}_k = 0 \text{ (for } m = 1, n) \qquad (20)$$

Minimizing (18) subject to these constraints gives

$$\underset{\sim}{w}_k = \underset{\sim}{d}_k + \sum_{m=1}^{n} \lambda_m \underset{\sim}{M}^{-1} \underset{\sim}{d}_l \qquad (21)$$

I now write $\underset{\sim}{\lambda}$ for the vector (dimension n) of multipliers λ_m, and $\underset{\sim}{D}$ for the N by n matrix whose m'th column is $\underset{\sim}{d}_l$. (20) and (21) become

$$\underline{D}^T \underline{w}_k = 0 \qquad (22)$$

$$\underline{w}_k = \underline{d}_k + \underline{M}^{-1} \underline{D} \underline{\lambda} \qquad (23)$$

The multipliers $\underline{\lambda}$ are given by multiplying (23) by $\underline{D}^T$ and using (22)

$$\underline{\lambda} = - (\underline{D}^T \underline{M}^{-1} \underline{D})^{-1} \underline{D}^T \underline{d}_k \qquad (24)$$

Substituting (22) and (23) in 18) gives

$$\langle (r_k - q_k)^2 \rangle = (\epsilon_k^o)^2 + 1 - \underline{w}_k^T \underline{w}_k \qquad (25)$$

It should be noted that this estimate of the interpolation error is arrived at assuming that the method, and all the estimated errors and correlations used, are perfect. In practice it was found that this occasionally gave unrealistically small values. Possible reasons for this are discussed in sections 3 and 4. To prevent these small values leading to the rejection of good data, an arbitrary additional error ϵ^m is added in the ECMWF system. A datum is thus considered to have failed the ckeck if

$$(r_k - q_k)^2 > T^2 (\langle (r_k - q_k)^2 \rangle + (\epsilon^m)^2) \qquad (26)$$

The tolerance T is currently assigned the value 4, and $(\epsilon^m)^2$ is 0.1.

Grid point analysis

In order to be able to use interpolation equation (12), while giving zero weight to data which have been included in M but subsequently rejected using (26), we need to minimize (7) subject to constraints like those in (22).

Manipulations like those of Section 3c give

$$\underline{w}_k = \underline{M}^{-1} \underline{p}_k - \underline{M}^{-1} \underline{D} (\underline{D}^T \underline{M}^{-1} \underline{D})^{-1} \underline{D}^T \underline{M}^{-1} \underline{p}_k \qquad (27)$$

Substituting (27) in (6) gives the equivalent of (11)

$$\underline{c} = \underline{M}^{-1} \underline{q} - \underline{M} \underline{D} (\underline{D}^T \underline{M}^{-1} \underline{D})^{-1} \underline{D}^T \underline{M}^{-1} \underline{q} \qquad (28)$$

As we might expect, the solution of the full set of equations constrained to give certain data zero weight is identical to that of a reduced set excluding those data. This reduced set has matrix

$$\underline{M}' = \underline{M} - \underline{M}\underline{D}\underline{D}^T - \underline{D}\underline{D}^T \underline{M} + \underline{D}\underline{D}^T \underline{M}\underline{D}\underline{D}^T \qquad (29)$$

where the additional terms leave out rows and columns l_m. Comparison of (27) and (28) with (8) and (11) shows that the matrix inverse $\underline{M}^{-1}$ has been replaced by

$$(\underline{M}^{-1})' = \underline{M}^{-1} - \underline{M}^{-1} \underline{D} (\underline{D}^T \underline{M}^{-1} \underline{D})^{-1} \underline{D}^T \underline{M}^{-1} \qquad (30)$$

Simple algebra yields

$$\mathbf{M}' \, (\mathbf{M}^{-1})' = \mathbf{I} - \mathbf{D}\mathbf{D}^{T} \qquad (31)$$

which is the identity matrix with the same rows and columns left out.

These equations are not as difficult to compute as first impressions might lead you to think, because of the sparce nature of $\underline{D}$ and the small number n. For instance for n = 1, l_1 = k, (30) becomes

$$(\mathbf{M}^{-1})' = [(M^{-1})_{i\,j} - (M^{-1})_{i\,k}\,(M^{-1})_{k\,j}/(M^{-1})_{k\,k}] \qquad (32)$$

There are several ways of reorganising these equations to perform the computations. One can calculate the modified matrix inverse using (30), or the modified weights using (23). Another reorganisation was used by Craven and Wahba (1979) for cross validation; this will be discussed in section 5.

The main advantage of using these equations is obtained for N large, n small, using the same basic matrix repeatedly. One algorithm for doing this is:-

a. Form a single matrix and inverse for all data.

b. Use (23) to (26) to check each datum in turn.

c. If any data fail reject the worst and use (30) to remove it permanently from the matrix inverse.

d. If more than one datum failed repeat from b.

e. If no data failed go on to do the analysis using (11) and (12).

This sequence makes it less likely that a bad datum will cause rejection of nearby good ones, and enables quality control and analysis to be combined.

These equations allow for correlated observational errors, but the check is only of the uncorrelated part. In interpolating the best estimate of an observed value using (21)-(24) any correlated part of the observation error is estimated as well. Thus observations from an observing system with abnormally large correlated errors pass the test. It is possible to derive checking equations to detect abnormally large correlated errors by replacing constraint (20) by one specifying that the analysis error should be uncorrelated with the observation error, ie $\langle \alpha_k \beta_k \rangle = 0$. This gives

$$(\mathbf{W}_k^{T} - \mathbf{b}_k^{T})\, \mathbf{W}_k = 0 \qquad (33)$$

Note that for uncorrelated errors this is equivalent to (20). The equations resulting from this constraint are however more complicated to evaluate, requiring separate knowledge of the observation and prediction error covariance components of M, rather than just their sum. Our knowledge of the correlation structure of observation errors does not justify such complication. It is probably better to check for correlated

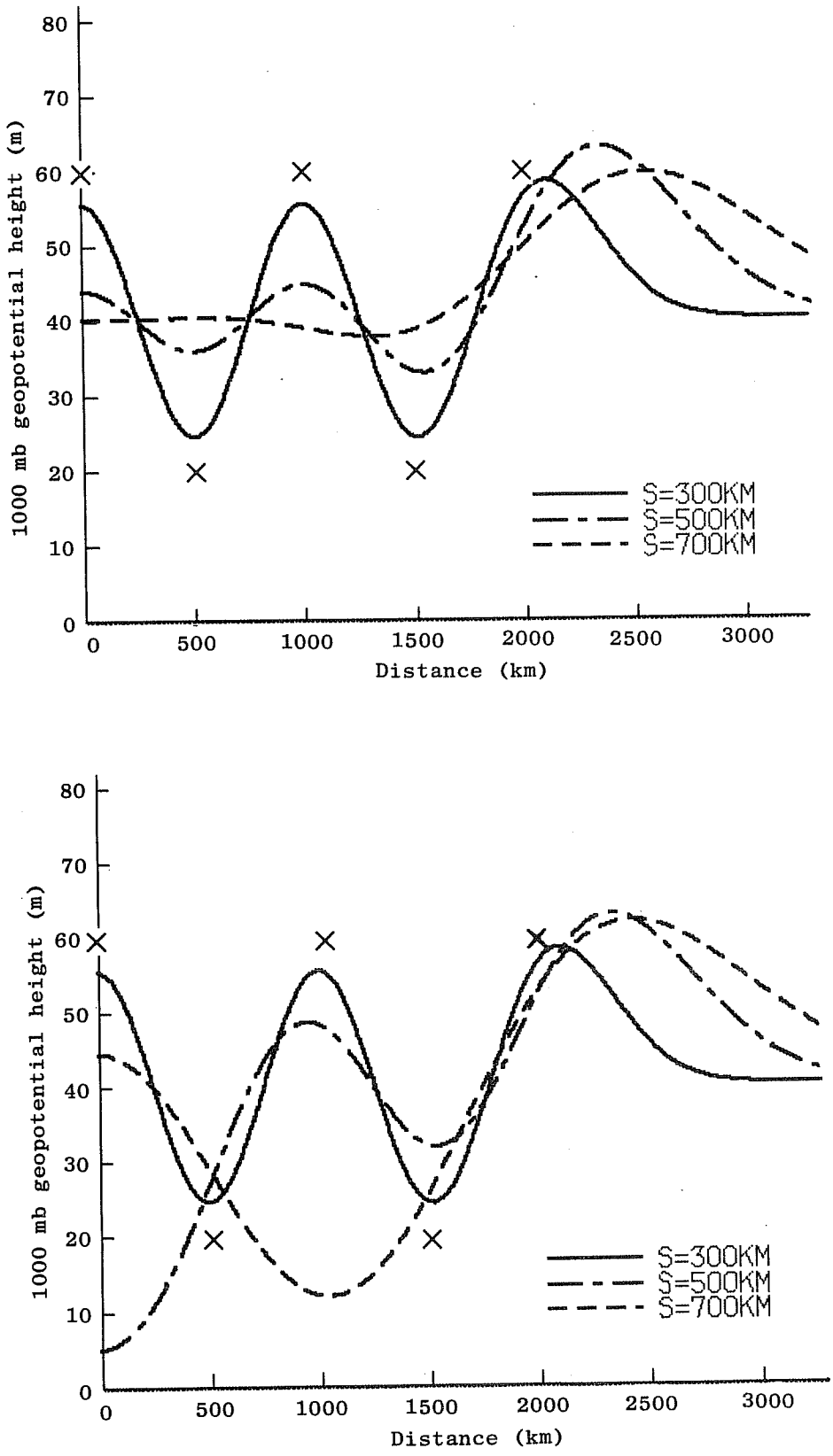
errors by removing all data from the observing system concerned using (20), checking them all using (26), and combining the results to either reject or accept the entire system.

One difficulty in partitioning $\underset{\sim}{M}$ into prediction and observational error components is that this split is dependent on the scale resolved by the analysis, because of the definition we have used for the truth T. This is in fact unnecessary for the basic quality control method based on (18) and (20). We can instead define the prediction error covariances and the observational error covariances in terms of deviations from the actual truth, and then partition the prediction error covariances at any spectral truncation limit. Resolved scales of this partitioned covariance function are what we have previously called the prediction error covariance (from a spectrally truncated 'truth'), and unresolved scales are what we have previously called observational errors of unrepresentativeness. Since the quality control equations only use the total matrix M, they are independent of this spectral truncation limit. Equations for this are derived in Appendix 3.

2.3 Examples

In order to be able to correctly quality control observations one must have redundant information. My first example (from Lorenc 1981) illustrates in an idealised case what can happen if this redundant information is not available. Figure 1 shows a 1000 km wavelength wave in the 1000 mb height field, just resolved by 9 observations with a 500 km spacing and deviations of alternately +20 m and -20 m from the background which is 40 m everywhere. The top half shows analyses of this wave for three different values of the prediction error scale parameter s. Only half of the symmetric situation is shown. The bottom half shows the same analyses after application of the quality control algorithm described in the last section. For small scales (s = 300 km) there is no redundant information, and the analysis is unchanged. For larger scales there is redundant information, but not enough to say clearly which of the data are wrong, so the quality control is not robust, and different scale specifications reject different data. Note that the analysis differences between the curves in the bottom half of Figure 1 are much larger than those in the top half. This is in agreement with the finding of Hollingsworth et al (1985) that large analysis differences are often associated with differences in quality control. However this example shows that they can be a symptom of insufficient data redundancy rather than poor quality control methods.

Since observation networks are not designed to give large amounts of redundant information, every effort should be made to use whatever information there is; this implies that multivariate analysis methods should be used for quality control. Other examples from Lorenc (1981) demonstrate this clearly. For instance temperature soundings give information about the geostrophic wind shear, and if sufficently accurate they can be used together with a low level wind to check an upper level wind. Except when all used together in this way the observations contain little redundant information. However the advantage to be gained by using data multivariately in this way decreases if the background field is of similar accuracy to the observations.



1. Analyses of a 1000 km wavelength wave from nine observations marked X for various horizontal prediction error correlation scales (s). Top: without quality control of data. Bottom: with quality control rejection of some data. Only half of the symmetric situation is shown.

RADIOSONDES

LEVEL		RMS 6HR FC		RMS ANAL.		MEAN 6HR FC		MEAN ANAL.	
MB	N	MO	EC	MO	EC	MO	EC	MO	EC
100	Z 509	60	50	62	38	32	17	44	7 M
250	Z 571	46	38	45	26	28	16	34	3 M
500	Z 590	30	21	26	14	18	7	17	-1 M
1000	Z 564	22	22	15	19	8	-4	2	-5 M
250	T 573	2.2	2.0	1.3	1.7	-.3	.2	.0	.1 K
500	T 591	1.7	1.7	1.3	1.4	.3	.2	.6	-.1 K
700	T 583	1.7	1.7	1.2	1.6	.4	-.0	.5	-.2 K
850	T 556	2.2	2.4	1.5	2.2	.3	-.3	.5	-.5 K
500	RH 579	25	29	18	25	-8	-16	-6	-14 %
700	RH 577	21	23	15	21	1	-10	0	-8 %
850	RH 553	20	22	13	19	2	-10	0	-6 %
SATEM (ref level 1000 mb)									
-100	DZ 617	39	37	30	34	16	8	14	4 M
-500	DZ 617	33	23	24	19	18	9	16	5 M

250 WIND

		VECTOR RSM				MEAN SPEED			
TEMP	532	8.4	7.5	4.8	5.3	.0	.6	.4	.4 M/S
T. SHIP	7	12.1	9.6	5.4	7.8	-.4	1.9	.3	2.6 M/S
PILOT	91	10.5	8.9	6.1	5.0	-.1	-.2	-.2	-.2 M/S
AIREP	464	11.2	9.9	8.7	9.1	.7	1.1	.0	.3 M/S
SATOB	51	9.9	8.8	7.2	6.0	-1.0	-1.6	-.1	-.9 M/S

1 Table showing mean and rms differences between observations valid at 12 GMT 7 June 1984 and the corresponding Met Office and ECMWF 6 hour forecast background fields and analyses.

Considerable effort has been put into tuning the ECMWF system to reduce the rms errors of the background field, by controlling the noise generated in the total analysis-initialisation-forecast cycle. This means that average background field errors are small, as shown in Table 1 (from Lorenc 1984), and statistics to this effect have been provided to the quality control algorithms (Hollingsworth 1984), which therefore use the background to provide much of the redundant information for quality control. Thus for the current ECMWF system use of a multivariate analysis method does not have much direct effect on quality control. In four recent test analyses the average number of rejected data was 76; when the multivariate coupling between height and wind was removed only on average 4 of these quality control decisions changed (D Shaw, personal communication). Based on the limited number of changes, which all occurred near the surface or in the upper stratosphere, D Shaw concluded that the multivariate check was acting beneficially. The observations near the surface, which could be better examined subjectively, were judged to be better evaluated with the multivariate check, which is modestly more stringent than a univariate one.

3. BAYESIAN APPROACH TO QUALITY CONTROL

I mentioned in section 2.1 the apparent problem that specifying an observation to be more accurate in an OI based scheme leads to it being rejected more often. Furthermore, in a scheme where all other parameters can at least in principle be related to statistics of model and observational errors, it should be possible to base the rejection criterion (set to 4 in section 2.2) on statistical evidence. A formalism for doing this, and removing the first problem, is provided by Bayes' Theorem.

This enables us to calculate the likelihood that any statement is true, based on prior estimates and new evidence. The OI equations can be derived using Bayes' Theorum and assuming normally distributed (Gaussian) errors. Let us consider the simplest case:-

T = the true value is t
O = the observed value is o
B = the prior (background) estimate is b

Then $P(B|T)$ is the probability that B is true given that T is true. For a normal background error distribution

$$P(B|T) = N(b-t, V_b)$$

where $$N(x,V) = (2\pi V)^{-1/2} \exp(-x^2/2V)$$

Then with no other information

$$P(T|B) = N(b-t, V_b)$$

From now on we shall understand knowledge of B always, and not represent it explicitly, eg we shall write $P(T)$ for $P(T|B)$.

Bayes' Theorum states that

$$P(T|O) = P(O|T)\,P(T)/P(O) \qquad (34)$$

$$= P(O|T)\,P(T)/\int P(O|T)\,P(T)\,dt$$

If $P(O|T) = N(o-t, V_o)$ then this gives us

$$P(T|O) = N(t-a, V_a)$$

where $1/V_a = 1/V_o + 1/V_b$, and $a = V_a\,(o/V_o + b/V_b)$.

This is shown in figure 2 for four different values of o. The values of a, the expected true value (the analysis value), and V_a are identical to those given by OI for this trivial zero dimensional case. A general Bayesian derivation of OI is given in Appendix 1.

Let us now extend this to a case where gross errors are possible. To proceed we need prior assumptions about the characteristics of these gross errors. Assume that the probability of a gross error occurring is independent of T, and that if a gross error has occurred the observed value is useless, also being independent of T, while if a gross error has not occurred the observation error is normally distributed as before:-

G = the observation has a gross error

$$P(G\Omega T) = P(G)\,P(T)$$

$$P(O|\bar{G}\Omega T) = N(o-t, V_o)$$

where $\bar{}$ denotes not, and Ω denotes and. Using Bayes' Theorum it is easy to show that:-

$$P(G|O) = P(G)/(P(G) + P(\bar{G})\,P(O|\bar{G})/P(O|G)) \qquad (35)$$

$$P(T|O) = P(G|O)\,P(T) + P(\bar{G}|O)\,P(T|O\Omega\bar{G}) \qquad (36)$$

$$P(O|\bar{G}) = N(o-b, V_o + V_b) \qquad (37)$$

We still have not completely specified P(O|G); it is assumed independent of T, but we need to know the distribution. Let us assume that over the region of interest (ie where P(O) and P(T) are significantly non-zero) it is a constant k. Climatologically unlikely observations would be rejected at an early stage, and for some sources of gross error (eg position errors) a climatological distribution is appropriate, so it is reasonable to assumed that the error variance V_g of gross errors is related to the climatological variance. For the flat distribution assumed, $k = (3/4V_g)^{1/2}$. An example showing P(T|O) for four values of o is shown in figure 3; this can be compared with the case without gross errors (figure 2). Note that for $(o-b)^2/(V_o + V_b)$ large then the analysis distribution with the possibility of gross errors becomes distinctly bimodal. Practically, since operationally we need a single 'best' analysis, we then have the problem of picking the 'best' value. One can argue for the mean, the median, or the mode; the variation of these with o-b is shown in Figure 4. If we take the mean then we can see that in OI terms this is equivalent

to reducing the weight given to the observation by the factor P(G|O), shown in Figure 5a. However for bimodal distributions the mean value can be rather unlikely to actually occur. Frequently in likelihood theory the modal value is taken. For bimodal distributions there is an abrupt jump in this value when one peak becomes larger than the other. For small a priori probability of gross errors this jump occurs approximately when

$$(o-b)^2 > (V_o + V_b) \ln((P(\bar{G})/P(G))^2 (V_g/V_o) /1.5\pi) \qquad (38)$$

If instead we calculate when $P(G|O) > 1/2$ then we get

$$(o-b)^2 > (V_o + V_b) \ln((P(\bar{G})/P(G))^2 (V_g/(V_o + V_b)) /1.5\pi) \qquad (39)$$

For small a priori probability of gross error the position of the rapid change in the median value is also approximately given by (39). The existence of the rapid changes in distribution parameters visible in figure 4 justify approximating the behaviour by simple acceptance or rejection, ie by saying that the 'best' value a is given by

$$a = \frac{o/V_o + b/V_b}{1/V_o + 1/V_b} \text{ for } (o-b)^2 < T^2 (V_o + V_b) \qquad (40)$$

$$a = b \text{ for } (o-b)^2 > T^2 (V_o + V_b)$$

(38) and (39) now give us some guidance on how to choose the tolerance T objectively. The criterion based on the mode (38) is only appropriate if we want to maximise the probability of being very nearly right; for small V_o we do this by choosing an analysis value near the observed value even when P(G|O) is large, since because of its larger variance the background is right less often. (39), which describes the behaviour of both the mean and (approximately) the median, is a better general criterion, and if T is based on this then (40) approximately gives the local mode nearest to the mean or median. Effectively we are approximating P(G|O) in (36) by 0 when $P(G|O) < 0.5$ and by 1 when $P(G|O) > 0.5$. The values of T implied by (39) for various values of $V_g/(V_o + V_b)$, and P(G) are plotted in figure 5b. The equivalent approximation for the variance of the analysis is

$$1/V_a = 1/V_o + 1/V_b \text{ for } (o-b)^2 < T^2 (V_o + V_b) \qquad (41)$$

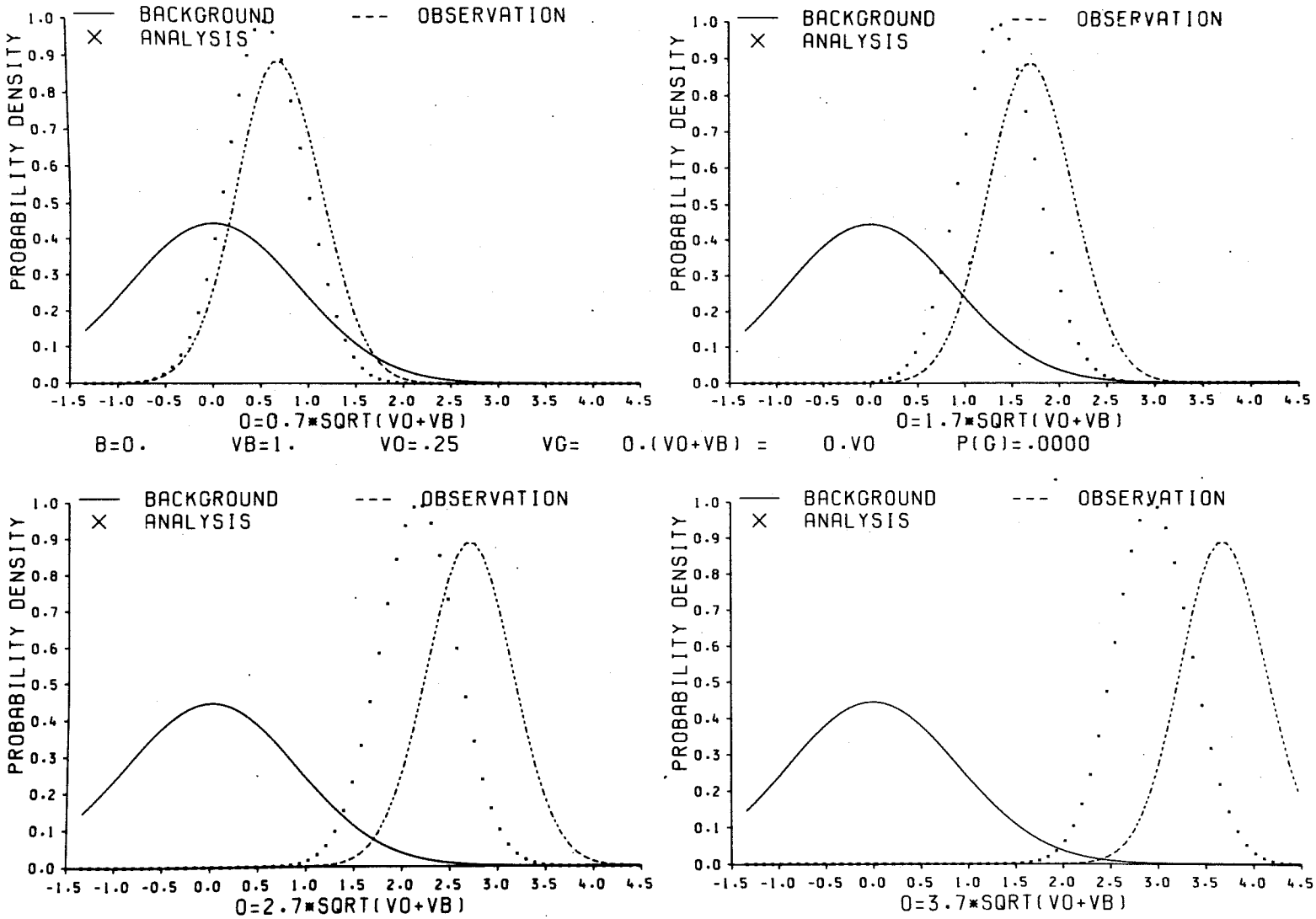
$$V_a = V_b \text{ for } (o-b)^2 > T^2 (V_o + V_b)$$

The actual variance of the posterior distribution given by (36) about its mean, and about the value given by (40), are shown in Figures 5c and d; they are much larger than given by (41). Thus although (40) gives a reasonable approximation to the 'best' analysis, (41) underestimates the actual analysis error. This has implications for the further use of the analysis for quality control of other data, and is one justification of ϵ^m in (26).

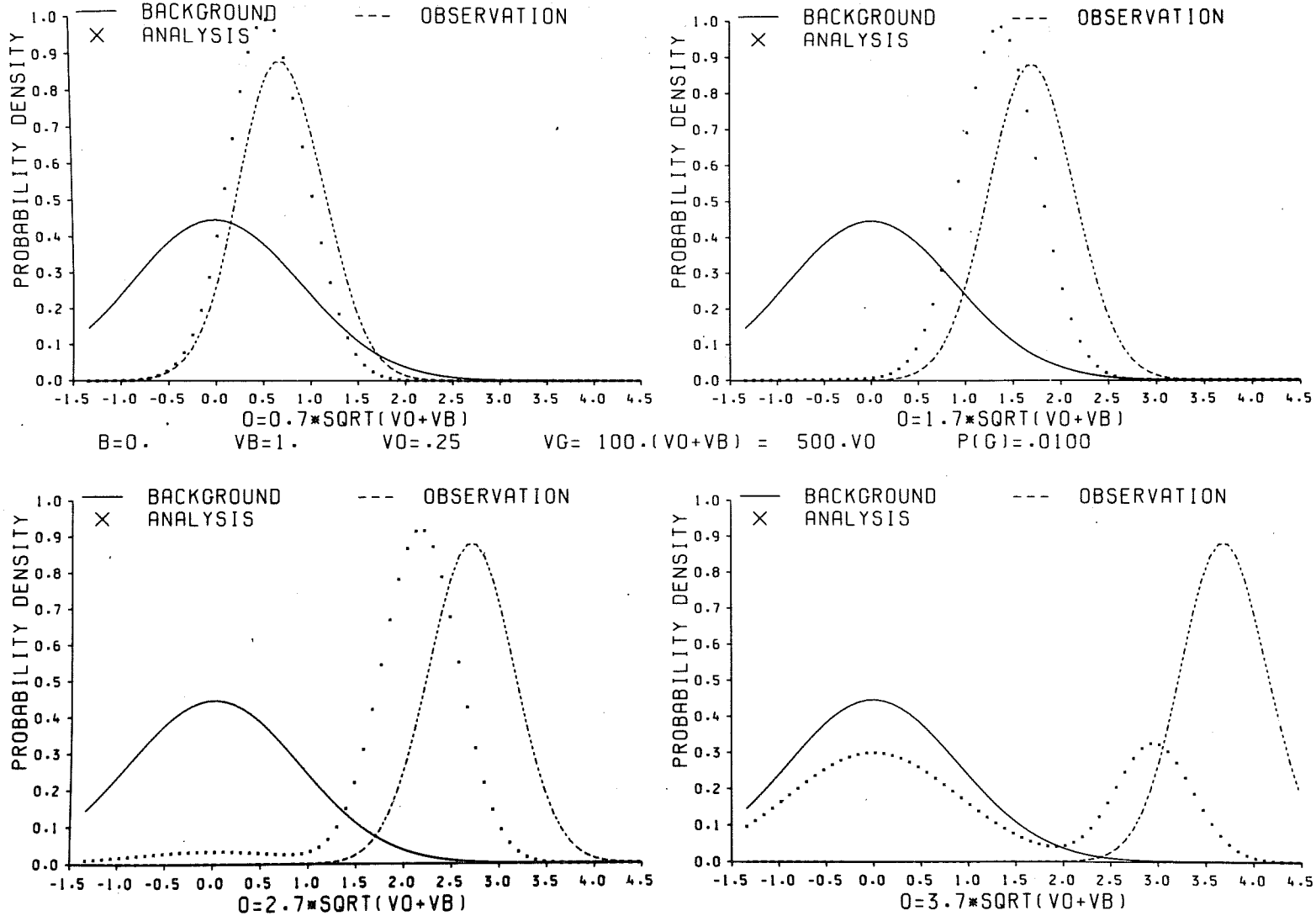
The Bayesian approach can be extended to two or more data each with uncorrelated gross errors possible. For instance for two observations it can be shown that

$$P(G_1|O_1 \cap O_2) = P(G_1|O_1)/B \qquad (42)$$

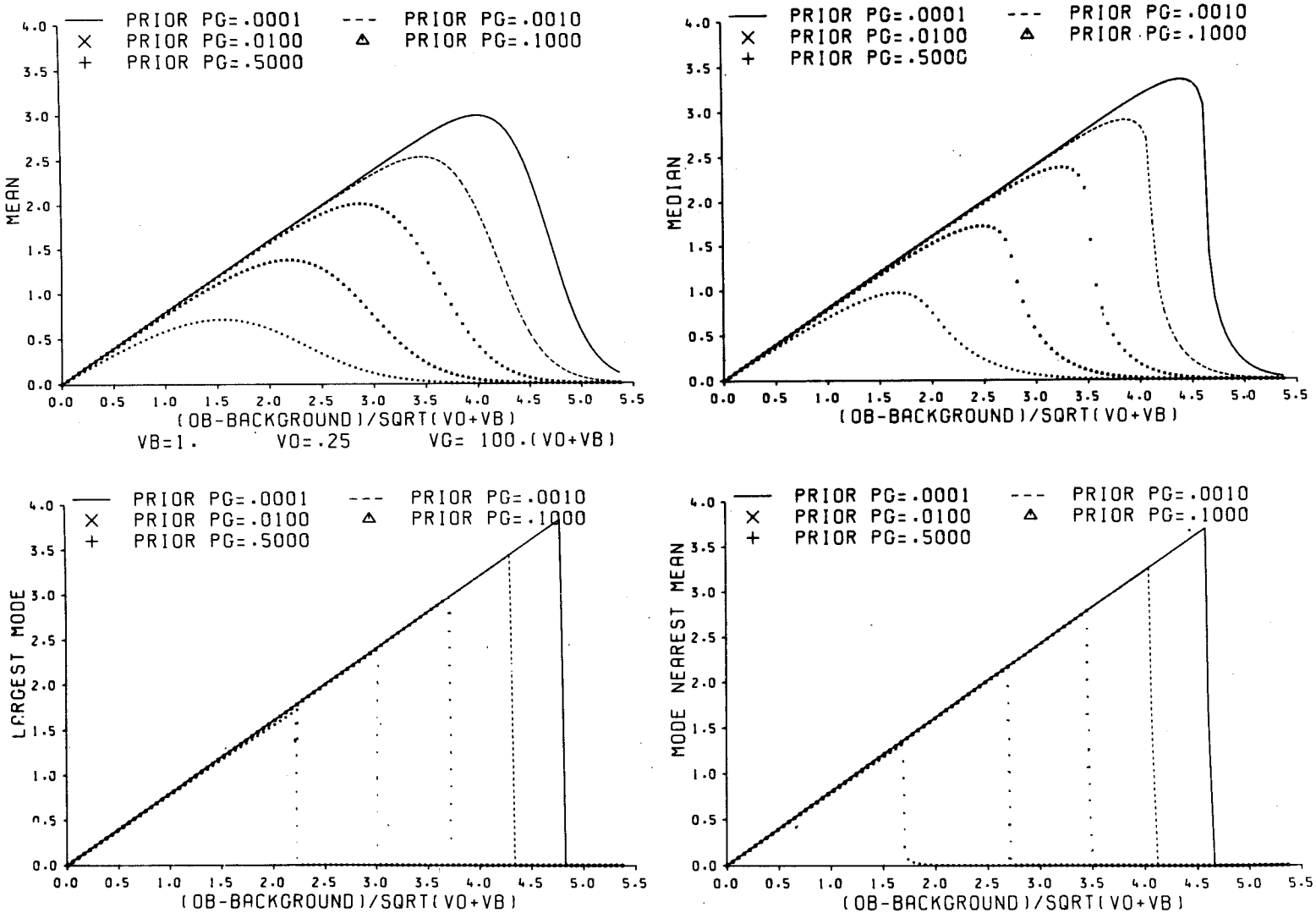
$$P(G_2|O_1 \cap O_2) = P(G_2|O_2)/B$$



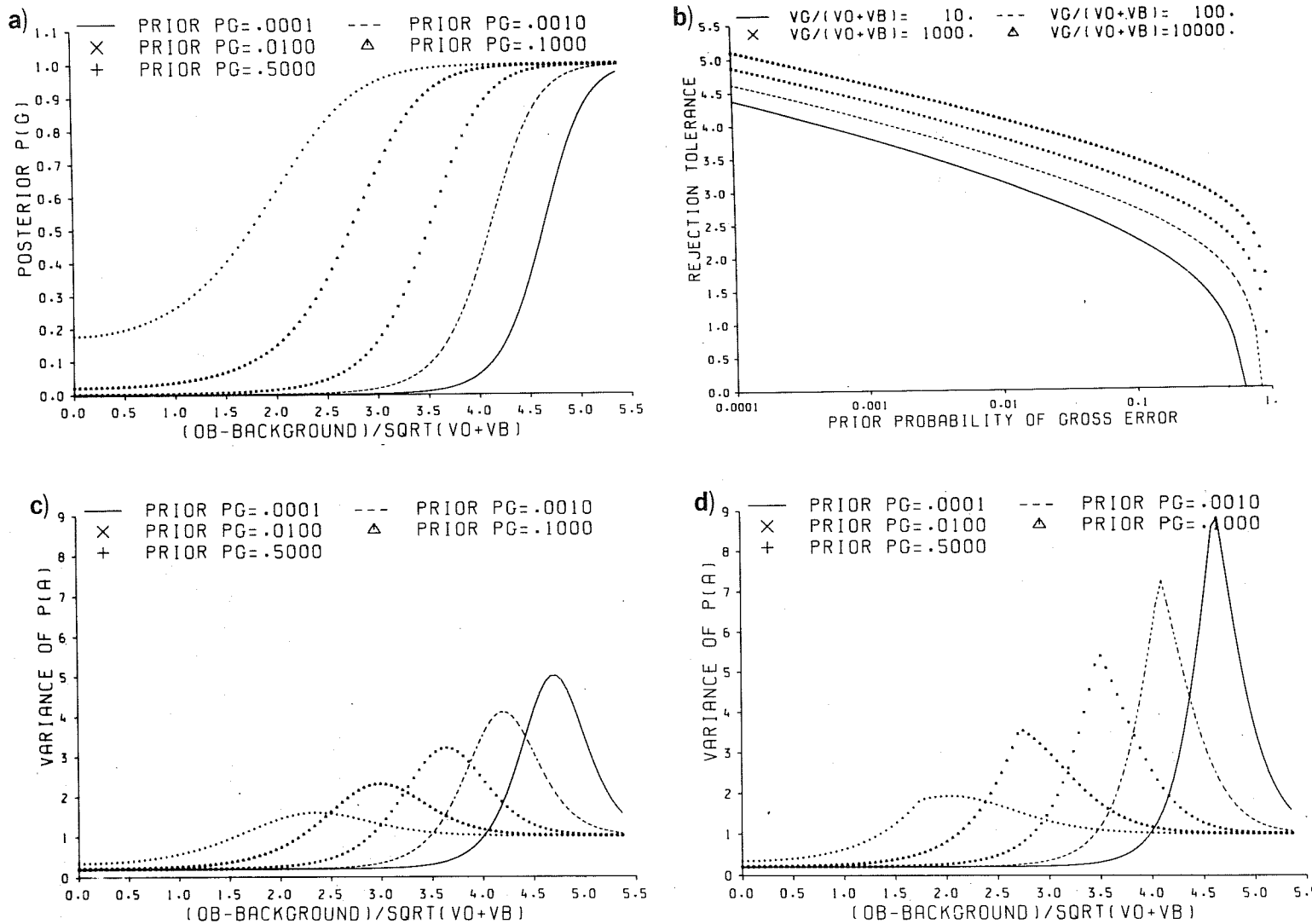
2. Probability density functions for background, observation, and Bayesian analysis, for four different values of o and a Gaussian observational error distribution.



3. As figure 2 for an observational error distribution equal to a Gaussian plus a small constant.



4. Mean, median, largest mode, and mode nearest the mean, of the analysis distribution, plotted against normalized observed minus background value, for various prior probabilities of gross error.



5. a: posterior probability of gross error, plotted against normalised observed minus background value, for various prior probabilities of gross error. b: Rejection tolerance T, based on the posterior probability of gross error being $1/2$, for various assumptions about the variance of gross errors, plotted against the prior probability of gross error. c: As a for the variance of the posterior distribution about its mean. d: as c for the variance about the value given by (40).

$$B = 1 - P(\bar{G}_1|O_1)\, P(\bar{G}_2|O_2)\, (1 - P(O_1 \cap O_2 | \bar{G}_1 \cap \bar{G}_2)/P(O_1|\bar{G}_1)P(O_2|\bar{G}_2))$$

$$= 1 - P(\bar{G}_1|O_1)\, P(\bar{G}_2|O_2)\, (1 - P(O_1|\bar{G}_1 \cap O_2 \cap \bar{G}_2)/P(O_1|\bar{G}_1))$$

some algebra yields a Gaussian expression for this last term:-

$$\frac{P(O_1|\bar{G}_1 \cap O_2 \cap \bar{G}_2)}{P(O_1|\bar{G}_1)} = \frac{(V_1 + V_b)(V_2 + V_b)}{V_1V_2 + V_1V_b + V_2V_b} \times \qquad (43)$$

$$\exp\left[\frac{-V_b^2}{2(V_1V_2 + V_1V_b + V_2V_b)} \left[\frac{(o_1-b)^2}{V_1+V_b} + \frac{(o_2-b)^2}{V_2+V_b} - \frac{2(o_1-b)(o_2-b)}{V_b} \right] \right]$$

It may be more practical to implement a sequential series of independent quality control checks on data, including 'buddy' checks between nearby data, particularly during any pre-selection of data for provision to the main analysis. These equations show how the Bayesian probability of gross error can be used to combine the results from these when individually they are inconclusive. Firstly a historical record of station reliability, checks on positions for ships, checks on format etc etc can be used to provide preliminary estimates of P(G). The observed values can then be compared with the forecast background using (35) to give P(G|O). Nearby pairs of data can then be intercompared and the probabilities updated using (42) and (43). The algebra for calculating an exact formula for $P(G_1|O_1 \cap O_2 \cap O_3)$ becomes complicated, but it is probably sufficiently accurate to use (42) and (43) recursively.

Of course a better estimate against which to check an observation can be obtained using multivariate OI and the equations of section 2.2. The algorithm proposed there can be expressed in the terminology of this section thus:-

b. Calculate $P(G_i|O_1 \cap O_2 \ldots\ldots O_N)$ for i = 1, N, assuming that $P(G_j) = 0$ for $j \neq i$.

c. If any data have this probability $> 1/2$ then assume that the datum with greatest value has P(G)=1.

d. Repeat from b.

Note that the assumption in b. that P(G) is either 0 or 1 for all data except that being checked enables us to use (35) and (39), replacing the background by the analysis, since the analysis will have Gaussian errors. The discussion of (40) and (41) shows that this analysis should be a reasonable approximation, but that its estimated error will be too small.

Although this method is better able to check for a single erroneous datum among many, it is not well able to cope with two or more bad data which happen to agree. This is discussed further in section 5.

4. COLLECTION OF STATISTICS

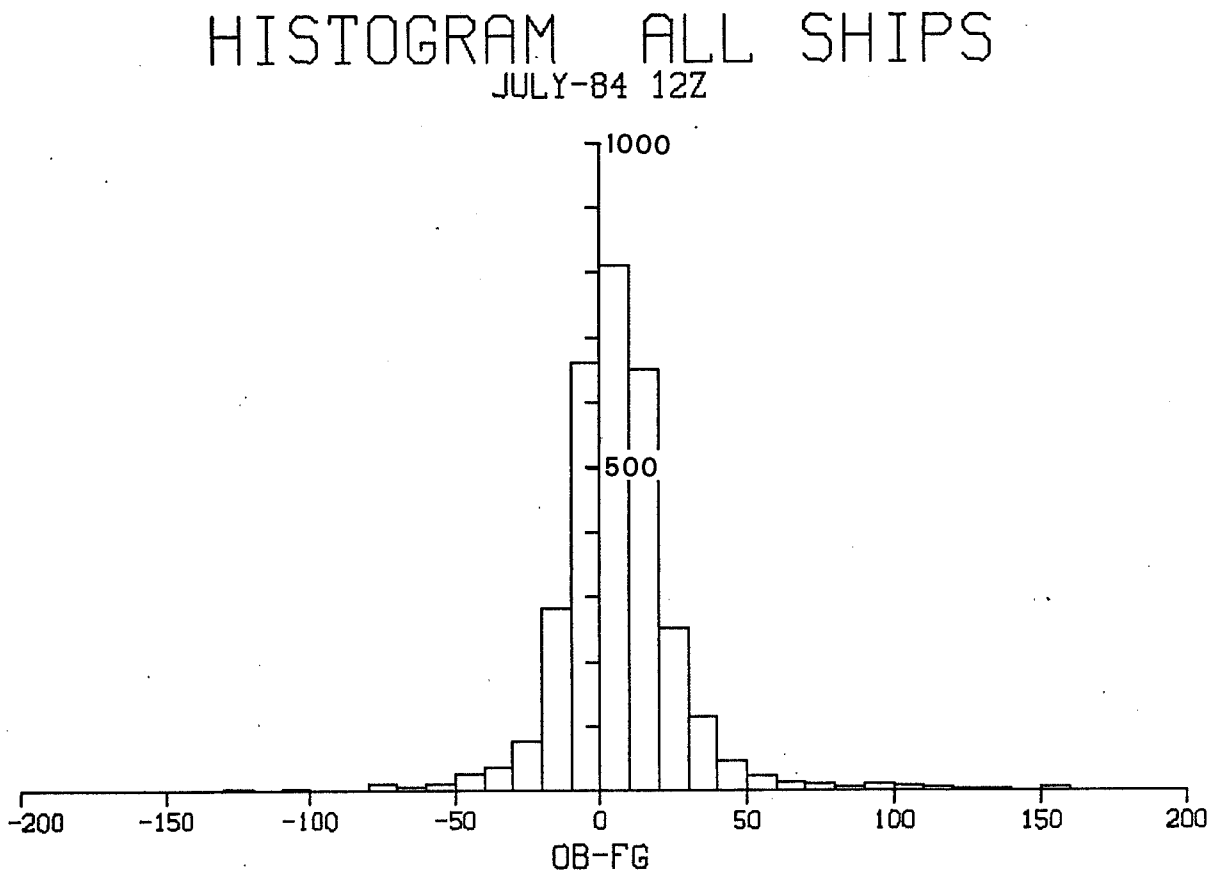
In order to apply equations like those in the last section one needs estimates of the frequency and a priori probability distribution of gross errors, as well as the estimated background errors and the observational errors for good observations. These statistics can be collected as part of

an operational system which is operating reasonably well, and used to improve it. An example is shown in Figures 6 and 7, which show histograms for the deviations between ship pressure observations (converted to 1000 mb height) and the background field for all observations and for rejected observations. The shape of Figure 7 can be used to justify assumptions about the distribution of gross errors, the proportion rejected (with some allowance for filling the gap near 0 of gross errors not detected) can be used to give a probability of gross error, and the variance of the histogram of accepted ships can be used to check $V_b + V_o$. The variance of a similar comparison of accepted observations with the analysed field should be $\leq V_o$ if the analyses method is working correctly. Such statistics collected during the analysis are dependent on preliminary estimates used in the analysis, and so some human monitoring and judgement is necessary to ensure that they converge towards reasonable values. As well as monitoring the quality of the observations the statistics can be used to monitor the performance of quality control, analysis, and forecast programs. For instance Figure 6 reveals an interesting bias, probably in the forecast pressure.

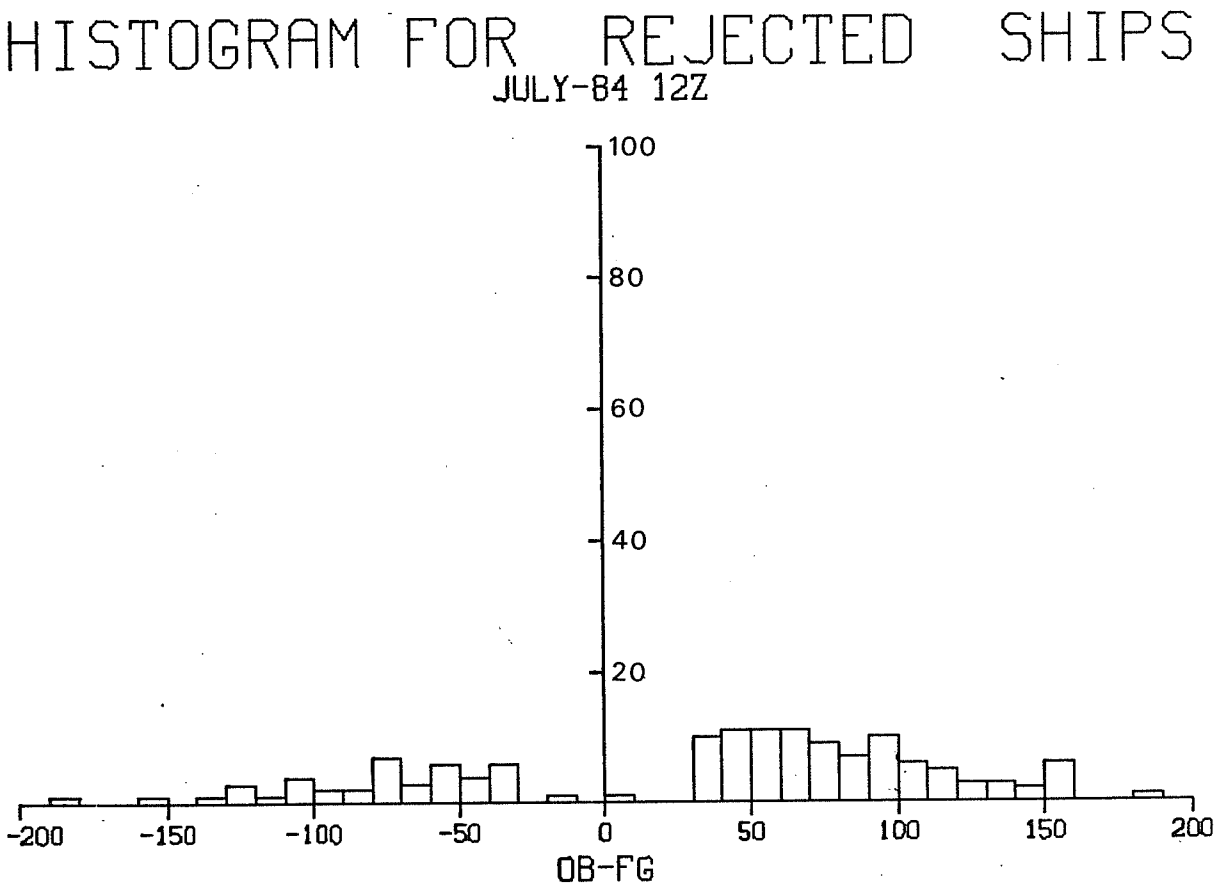
The collection of statistics is even more critical for quality control than for OI for several reasons: (1) statistics are needed for the distribution of rather rare events, (2) we need to know observation error E^o and prediction error E^p independently, rather than just their ratio, (3) the use of inappropriate prediction error correlations has much more effect on the estimated analysis error which is used in quality control than on the actual analysis error; for instance Franke (1984) showed that too large an estimate for the horizontal scale of prediction error correlations caused ϵ^a to be underestimated, although actual errors in the analysis were slightly increased. This is another justification for the term ϵ^m in (26).

It is precisely those observations which deviate from the background field and which are correct which contain the most new information, so we must take care not to reject them by underestimating the analysis error. This can occur if the estimated background error is too small, or if the prediction error correlations are too large. If we use time average statistics for these, independent of meteorological situation, these are both likely to happen in regions of developing small systems, near active jet streams. Both the resolved prediction error and the unresolved error of representativeness are likely to be larger than average in such areas, while the scale of prediction error is likely to be smaller than average. Thus there is a strong case for attempting to stratify the statistics used for quality control to take account of the actual meteorological situation. In the empirically tuned quality control algorithms used in the old Met Office 10-level model analysis scheme an objective method of recognising such active situations was successfully used.

The collection, maintenance and use of such a statistical database should be an integral part of an operational analysis system. it must be flexible; a great variety of types of error can be postulated, both systematic and random, in both the observations and the various components of the analysis system. It should be possible to use the database to investigate the significance of any type of error, before designing corrections or quality control tests to cope with it automatically.



6. Histogram of observed - background differences for all ship observations of sea level pressure (converted to 1000 mb height) from the ECMWF analysis system.



NR.OBS= 138.0

MEAN= 33.0 STD= 80.9

7. As figure 6 for ships rejected by the analysis quality control.

5. ALTERNATIVE METHODS

The Bayesian approach used in section 3 is like that of Purser (1984), but the result, a criterion for completely accepting or rejecting a datum, is not. This is because of the assumptions made about the random nature of gross errors, with bad data giving no information at all, lead to the bimodal posterior probability distribution which can be modelled by simple rejection or acceptance. For many types of observations, in particular for indirect remote sensing data, these assumptions about the characteristics of gross errors are inappropriate; a continuous range of varying degrees of error occurring. The equations of section 3 can be easily adapted to apply to the case where observations which are 'grossly' in error in fact still carry some useful information, by assuming that $P(O|G\Omega T) = N(o-t, V_g)$. For V_g large, as in figures 3, 4 and 5, results are little altered, but for smaller V_g and larger P(G) the posterior distribution has much less of distinct bimodal structure (figure 8), and it is not appropriate to approximate its parameters (figure 9) by equations like (40) and (41). Purser (1984) shows how the most likely analysis (ie the mode of the posterior distribution) for such a non-Gaussian distribution can be obtained by using the OI equations with an effective observation error given by

$$V_O = - \frac{\partial}{\partial a} \ln [P_O (o-a)]/(o-a) \qquad (43)$$

where P_O is the observational error probability distribution. Note that this gives the same analysis equation as before for $P_O(o-a) = N(o-a, V_O)$, but that for all other non-Gaussian distributions V_O is a function of a and the analysis equation is implicit. By iteration any reasonably behaved observational error distribution can be allowed for, if it is known. For distributions with longer tails than a Gaussian the effect is to increase V_O and decrease the weight for observations which deviate from the analysis. (The rejection criterion of section 3 is a limiting case of this). Other approximations can be suggested, either to the OI equations to make the iterations less expensive, or by using the deviation from the background as an estimate of o-a.

The methods described so far all require prior estimates of the statistics, although we did discuss in section 4 how statistics could be accumulated during the analysis to improve the estimates in subsequent analyses. A method which short-circuits this, and uses each analysis to estimate its own appropriate statistics, is called generalized cross validation (GCV). It was presented by Craven and Wahba (1979), and Wahba and Wendelberger (1980) for a variational spline-fitting analysis method, but it is suitable for OI since this can be couched in variational form. The basis of the method is to choose a few parameters in the statistics to minimize a function (GCVF) equal to a weighted sum of squares of the deviation of each observation from an analysis made not using it. Now these deviations are precisely what I have used to quality control the data, and so not surprisingly the matrix manipulations proposed by Craven and Wahba (1979) for efficient calculation of the GCVF are equivalent to those used in section 2 for quality control, if we assume that observation errors are uncorrelated. Extending my notation of section 2 so that r_k is

the standard OI analysis at the position of datum k, and r_k' is the analysed value not using datum k, then as long as datum k does not have correlated errors:

$$r_k' = (r_k - w_{kk}\, q_k)/(1-w_{kk}) \qquad (44)$$

Then GCVF, which is a weighted sum of $(r_k' - q_k)^2$, can be efficiently calculated by

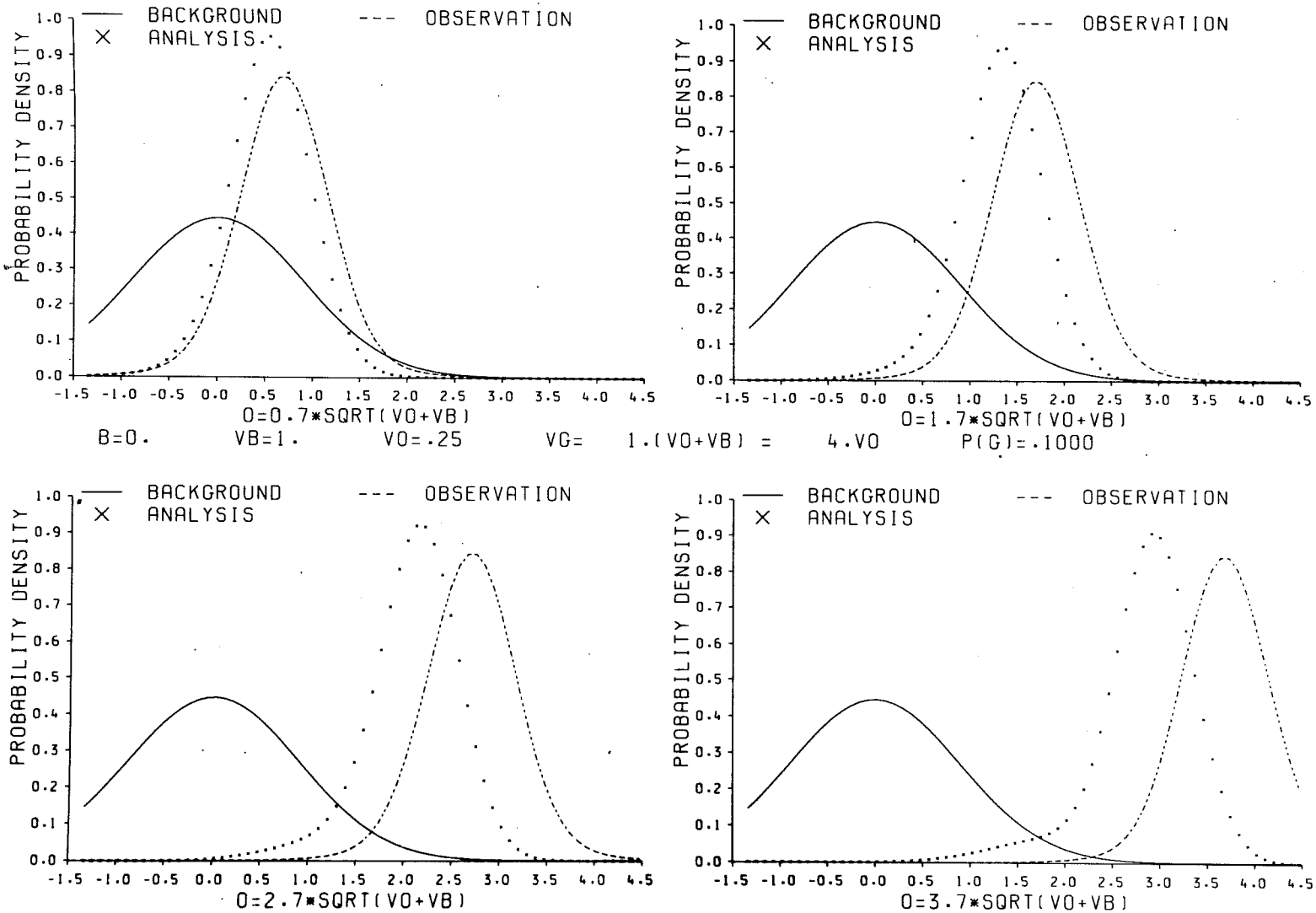
$$GCVF = \frac{1}{N}\sum_k (r_k - q_k)^2 (\epsilon_k^o)^{-2}/(1 + \frac{1}{N}\sum_k w_{kk}) \qquad (45)$$

Just like quality control, GCV is reliant on there being redundant information in the data; it could not cope with my example of figure 1. Generally only a few parameters in the statistics can be reliably calculated; usually just the ratio of observation error to prediction error. Quality control and GCV can be looked on as alternatives; in quality control prior statistics and redundant information are used to reject bad data, in GCV identical comparisons are used to produce revised statistics.

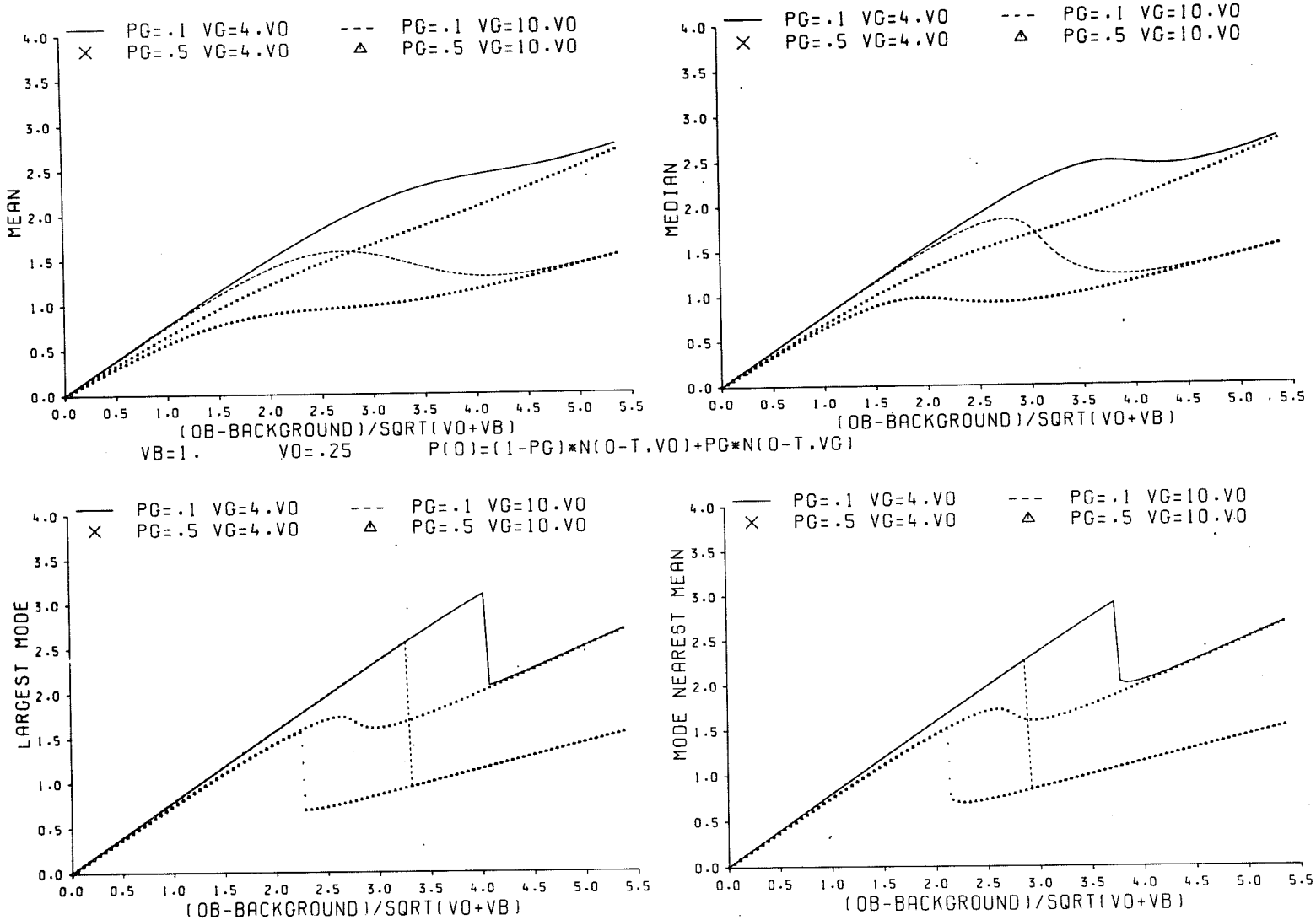
One method used in other disciples for fitting data while simultaneously eradicating wild points is the use of the L_1 norm (Barrodale 1968). This replaces the squared penalty function on the fit to observations (the first term in (A9) in Appendix 1) by a mean of absolute deviations. In zero dimensions this is equivalent to finding the median rather than the mean of the data; the median is independent of the values of extreme data. Such a technique might have applications to fitting data with errors whose characteristics are not well known, such as satellite temperature soundings. However it does not have as convenient mathematical properties and is not as easy to implement as the standard L_2 norm.

For some indirect observing systems, meteorological parameters which are not well observed affect the accuracy of the observations; one example is in the assignment of cloud motion winds to a particular height, another is in the retrieval of temperature profiles from observed radiances. Errors in the latter are particularly difficult to cope with; they are usually correlated and occasionally large. One bad example was studied by Adams (1984). An active trough over the UK had a very low tropopause behind it; the statistical retrieval algorithm used could not deduce this, and produced soundings with a climatologically appropriate tropopause. These therefore deviated from the background in a systematic way, as shown in figure 10. Other observations did agree with the background; the use of the satellite data as they were would degrade the analysis in this case. The main hope for correcting errors such as these is in a closer integration of the observation processing and analysis, using background fields or preliminary analyses to provide missing information such as tropopause height, or vertical structure for cloud motion wind level assignment.

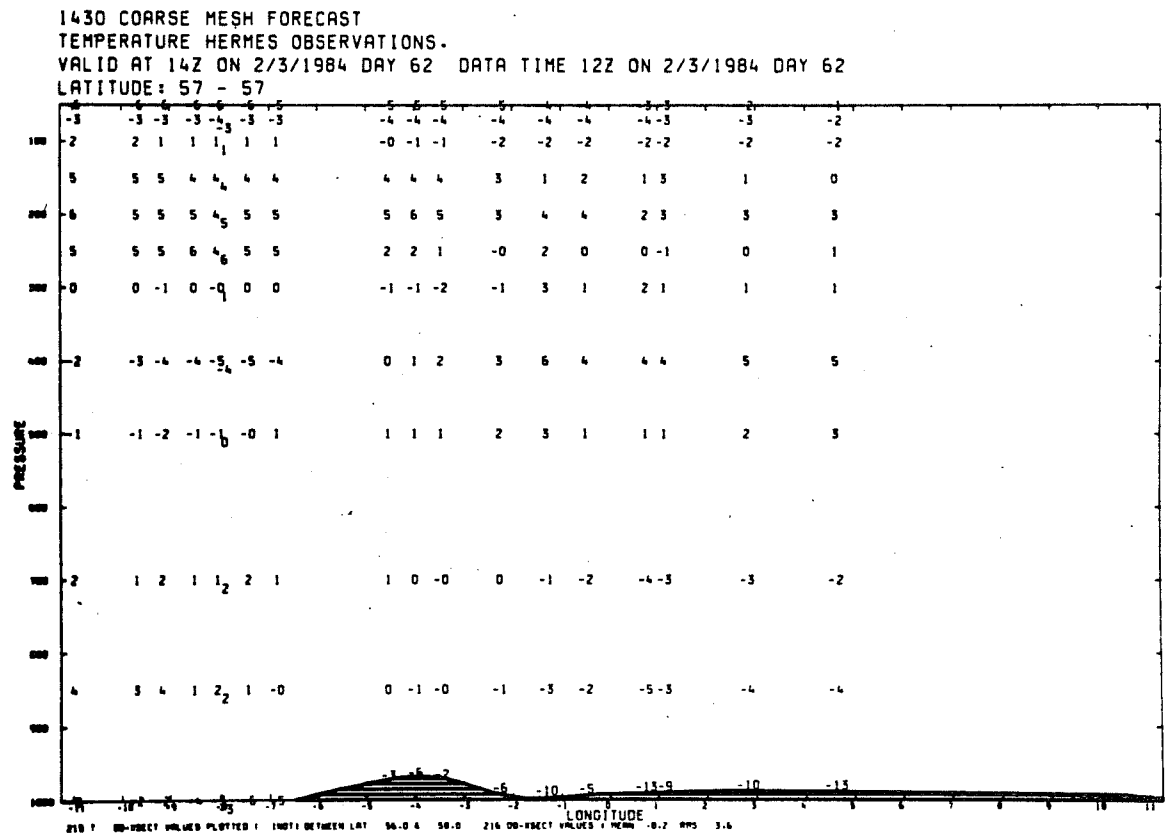
Sometimes entire observing systems can be grossly in error. Examples I have seen are:- erroneous orbit parameters causing all temperature soundings to be misplaced, high level cloud motion winds all assigned to low level, a radar misaligned so that all radiosonde wind directions were wrong, a station altitude wrong so that all radiosonde height data were



8. As figure 2, for observational error distribution equal to the sum of two Gaussians.



9. As figure 4, for observational error distribution equal to the sum of two Gaussians.



10. Cross section showing deviations of high resolution satellite temperature retrievals from a background field. The section runs from 11°W to 11°E and the deviations of all data near 57°N are shown in K.

wrong, and anomalous echos on a radar so that radar rainfall measurements were wrong. Such errors cannot be reliably detected using the equations of section 3, which rely on P(G) being independent between observations. Specific tests on particular observing systems can be incorporated in the OI checking procedure of section 2, by omitting and checking entire observing systems simultaneously, but no general expression for a rejection criterion can be given, since this depends on the likely nature and structure of gross error. Thus specific coding is necessary for each type of gross error, and for a long time to come we will have to rely on human monitoring to look for patterns in unforeseen errors and to form and test hypotheses as to their cause. Automatic analysis procedures can greatly aid this procedure however, by identifying areas where something appears abnormal, for closer study by the monitor. One useful tool for this is a display of the analysis increments (ie the difference from the background field). If those are abnormally large it is likely that either the forecast, or the observations, or the analysis process is in error.

The automatic methods of sections 2 and 3, if provided with reasonable statistics, should be able to cope with random gross errors at least as well as a human monitor, and because of their reliability and speed they are to be preferred. Human monitoring should be confined to areas where there is a distinct advantage over automatic processing; these are pattern recognition, and hypothesis formation and testing. One area where human skill at pattern recognition might be used to improve current automatic systems without the effort of scanning individual observations is in the specification of regions where the background field is likely to be abnormally erroneous. As automatic methods for this, and for detecting various types of correlated errors are developed, their performance will need careful monitoring. Human resources are probably better deployed doing this than in real time quality control of individual observations.

6. CONCLUDING REMARKS

I have shown that a statistically sound basis for quality control of observations can be provided if prior assumptions about the distribution and probability of errors are made. This implies that one needs to know the type of error one is looking for; there is no such thing as a general check for all errors. In particular different checks are needed for a single datum and for related errors in several data from one observing system.

In section 3 a Bayesian approach was used to derive equations for calculating the probability of gross error, assuming that grossly erroneous observations carry no information. it was argued that this should provide a criterion for accepting or rejecting the datum completely, as is done in most current analysis systems. It also provided a basis for combining consecutive, individually inconclusive, independent tests to give a final more conclusive criterion.

On the other hand for many indirect observing systems such as satellite temperature soundings we wish to allow for abnormally large errors in data which still contain some useful information. In this case complete rejection is inappropriate, and adaptive analysis methods should

be used such as a non-Gaussian Bayesian analysis if observational error distributions can be specified, or cross validation if they need to be estimated from the data. These were discussed in section 5.

Whatever the method used, lack of complete information redundancy means that results will be probabilistic in nature, based on statistical criteria. The accumulation of the relevant statistics is a crucial part of the quality control programs, as indeed it is for any statistically based analysis method, and should be considered at the design stage of any analysis suite. It should be possible to build up statistics of the probability and distribution of each type of error, and to use these to improve quality control criteria. This was discussed in section 4.

There remain several areas which are difficult to automate, and where human monitoring should be useful. These use the human skills of pattern recognition, hypothesis formation and testing, and flexibility in unforeseen circumstances. Unforeseen is a key word in this; human resources are best deployed after automatic methods of quality control, so that as these are extended to cope with more types of error, the human can concentrate on those areas where his pattern recognition skill can use cloud pictures or knowledge of meteorological structures to decide on borderline data. Eventually I believe that real-time routine quality-control of observations by a human will in most cases be found to be counter-productive, because the time taken delays forecasts, on not cost-effective, because of the skilled manpower needed to cope with the few cases where automatic methods can be bettered. However as automatic methods become more complex there will be an increasing need to monitor their performance on a non-real-time basis.

The programming and maintenance of the flexible, integrated, quality-control statistical-database and monitoring system I advocate would be difficult. The steady advance in computer technology should be used to produce a flexible modular high-level system, capable of expansion to cope with the multiplicity of specialised statistics and checks which accumulate to deal with specific types of error. Research in artificial intelligence, and self-teaching expert systems, should be used to aid the design of a system which, if the investment in software is to be justified, will have to outlast the current generation of computers.

Acknowledgements

Thanks are due to Lodovika Illari and David Shaw for providing histograms and the results of research with the ECMWF quality control system.

Appendices

1. A Bayesian derivation of OI

(This derivation was the notation of Purser (1984)).

Let the prior background field have a state vector B. (The elements of this can be grid points/variable values, but they can also be spectral modes). Deviations of this from the truth are represented by B' = B-T.

Then the background field error covariance matrix is given by $C = \langle B' B'^T \rangle$ (A1)

This has eigenvector matrix V and diagonal eigenvalue matrix λ which can transform the co-ordinates of B' to an orthogonal set:-

$$\langle (V^T B')\,(V^T B')^T \rangle = V^T C V = \lambda \qquad (A2)$$

Assume that in these orthogonal co-ordinates each element i is independently
normally distributed with variance λ_{ii}. Then

$$P(B) \alpha \exp - {}^1/_2\, (V^T B')^T \lambda^{-1} (V^T B')$$

$$\alpha \exp {}^1/_2\, B'^T C^{-1} B' \qquad (A3)$$

Similarly consider a vector of observed values O with error covariance E. Then

$$P(O) \alpha \exp -{}^1/_2\, O'^T E^{-1} O' \qquad (A4)$$

Now the elements of O and B are not the same. Let us assume the existence of a linear transformation D which converts from the basic state space of B to the observed parameters of O. If B is expressed as grid point values then the elements of D are interpolation coefficients, if B is in spectral coefficients then they are modes evaluated at the observation positions. If the observed parameter is an area average then D should reflect this. If the observed parameter is not one of the independent variables chosen for B (eg it might be a radiance) then D represents the appropriate transformation. Note that D need not be linear as long as it can be approximated linearly in the region of B and T. In its linear form D is a rectangular matrix. Using this we can always express T in the state space of B eg

$$O' = O - DT \qquad (A5)$$

Now Bayes' Theorum states that the probability of state A given B and O is given by

$$P(A \mid B \,\Omega\, O) \alpha P(O \mid B \,\Omega\, A)\ P(A \mid B) \qquad (A6)$$

Now if background and observation errors are uncorrelated then
$P(O \mid B \,\Omega\, A) = P(O \mid A)$

Also, without the knowledge of the observations, Bayes' Theorum gives us

$$P(A \mid B) = P(B \mid A)$$

Therefore (A6) becomes

$$P(A \mid B \cap O) \alpha \; P(O \mid A) \; P(B \mid A) \tag{A7}$$

$$P(A \mid B \cap O) \alpha \; \exp(-1/2 \; (O-DA)^T E^{-1} (O-DA)) \exp(-1/2 \; (B-A)^T C^{-1} (B-A) \tag{A8}$$

The maximum likelihood estimate of A (ie the mode of the posterior distribution) is given when this is a maximum. ln(x) is maximum when x is maximum, so this is given by

$$0 = \frac{\partial}{\partial A} [1/2 \; (O-DA)^T E^{-1} (O-DA) + 1/2 \; (B-A)^T C^{-1} (B-A)] \tag{A9}$$

where $\partial/\partial A$ is notation for a set of equations for each element of A. This gives

$$0 = C^{-1} (B-A) + D^T E^{-1} (O-DA)$$

$$= C^{-1} (B-A) + D^T E^{-1} (O-DB) + D^T E^{-1} D (B-A) \tag{A10}$$

If D is only a linear approximation to a non-linear function D' then

$$D'(B) = D'(A) + D (B-A) \tag{A11}$$

This gives

$$(A-B) = (D^T E^{-1} D + C^{-1})^{-1} D^T E^{-1} (O-D'(B)) \tag{A12}$$

This equation is an 'inside out' OI, which might be useful when the number of observations exceeds the order of B, eg for the analysis of a few planetary modes when B is in spectral form, since the matrix inverse required is of the order of B. To turn it the 'right way out' we need

$$(DCD^T + E)^{-1} (DCD^T + E) = I \tag{A13}$$

Then (A12) gives

$$(A-B) = (D^T E^{-1})^{-1} D^T E^{-1} (DCD^T + E) (DCD^T + E)^{-1} (O-D'(B))$$

$$= (D^T E^{-1} D + C^{-1})^{-1} (D^T E^{-1} DCD^T + C^{-1} CD^T)(DCD^T + E)^{-1} (O-D'(B))$$

$$A-B = (DC)^T (DCD^T + E)^{-1} (O-D'(B)) \tag{A14}$$

which is the standard OI equation.

2. The equivalence of OI and variational methods

Note that other variational approaches can also give (A9), the first term being a penalty function for the fit to the data, the second (particularly in spectral space if C is nearly diagonal) being a penalty on smoothness. The equivalence of the approaches is discussed by Wahba and Wendelberger

(1980). Note that for smoothing splines there is no penalty on constant or linearly varying fields. It can be seen from (A9) that this is equivalent to an infinite background error variance for these, ie they are totally derived from the data in the analysis. This explains the poor behaviour of the spline method when extrapolating beyond the edge of a data area.

3. The effect of spectral truncation

In the above derivation we have not considered resolution. Since we have derived generalized OI equations which are valid for B expressed in spectral modes, it is easy to include practical limits on truncation. Unlike in section 2.2, here we do not define T to be spectrally truncated. Instead we postulate the existance of a spectral truncation operator S with properties:-

$$SS = S = S^T \tag{A15}$$

$$SB = B \tag{A16}$$

$$S \langle T \rangle = \langle ST \rangle \tag{A17}$$

(A17) states that the best estimate of the truncated analysis (as derived in section2.2), is given by the spectral truncation of the full analysis (as derived in the appendix). This might not be true if there are significant correlations across the truncation limit, for example if the truncation does not resolve a fjord. Using these (A14) gives

$$SA = B + (DSC)^T (DSCD^T + D(I-S)CD^T + E)^{-1} (O-D'(B)) \tag{A18}$$

Here SC is the truncated covariance function used in section 2.2 and $D(I-S)CD^T$ the part of observational errors due to unrepresentativeness. Note that for certain types of area average observation we have said that D has smoothing properties; for these the second term will be small because D(I-S) is small.

References

Adams, Wendy 1984 Assessment of HERMES data: A case study comparison with the operational analysis for 2nd March 1984. Met O 11 Tech Note 195.

Barrodale, I 1968 L_1 approximations and the analysis of data. Applied Stat. 17 51-57.

Craven, P and Wahba, G 1979 Smoothing noisy data with spline functions - estimating the correct degree of smoothing by the method of generalised cross-validation. Numer Math 31, 377-403.

Franke, R 1984 Sources of error in objective analysis. Tech. Report NPS-53-84-0003 Naval Postgraduate School, Monterey CA. Submitted to Mon. Wea. Rev.

Hollingsworth, A	1984	Recent developments in data assimilation at ECMWF. ECMWF/WMO Seminar on 'Data assimilation systems and observing system experiments with particular emphasis on FGGE'.
Hollingsworth, A Lorenc, A C, Tracton, M S, Arpe, K, Cats, G, Uppala, S, Kallberg, P	1985	'The response of Numerical Weather Prediction Systems to FGGE IIb Data, Part I: Analysis. Quart. J. R. Met. Soc. 111 (to appear).
Lorenc, A C	1981	A global three-dimensional multivariate statistical analysis scheme. Mon. Wea. Rev. 109 701-721.
Lorenc, A C	1984	Data assimilation by repeated insertion into a forecast model - principles, practice, problems and plans. ECMWF Seminar on data assimilation systems and OSE with particular emphasis on FGGE. September 1984. Met O 11 Tech Note 193.
Purser, R J	1984	A new approach to the optimal assimilation of meteorological data by interative Bayesian analysis. Preprints, 10th conference on weather forecasting and analysis. Am. Met. Soc. 102-105.
Wahba, G and Wendelberger, J	1980	Some new mathematical methods for variational objective analysis using splines and cross validation. Mon. Wea. Rev. 108 1122-1143.